**Global\_Mobility\_Report**

* + metro\_area: can be ignored since it only empty data and Kabul Metropolitan area
  + iso\_3166\_2\_code: can be ignored since we can use country code feature column instead
  + census\_fips\_code: can be ignored since it has no data
  + sub\_region\_2: can be ignored because they are the sub regions under “sub\_region\_1”. So we can simply use “sub\_region\_1”
  + country\_region\_code: can be ignored because we can instead use “country\_region”

Analysis:

* Average or % change from baseline per country on year 2020 and 2021
* Comparison of impact between different sectors regardless of country
* Comparison of impact (regardless of sector) between countries
* Country Heat map showing average impact (from all sectors)
* Word Heat map showing average impact (from all sub regions)

Prediction:

**COVID-19\_cases\_TX**

* + State and state\_fips\_code: can be ignored since the data is only for Texas
  + county\_name: can be ignored and just use “state\_fips\_code” (Federal Information Processing Standards code) since it’s just codes for counties in US

Analysis:

* Comparison of confirmed cases and deaths between 2020 and 2021
* Show count of confirmed cases and deaths in each county

Prediction:

**COVID-19\_cases\_plus\_census**

* county\_name: not needed since we already have “county\_fips\_code”
* state: not needed since we already have “state\_fips\_code”
* date: not needed since the data is only for one date (1/19/2021)
* geo\_id: not needed since we already have “county\_fips\_code”
* median\_year\_structure\_built: not relevant to our data since it’s about housing
* ren\_burden\_not\_computed: not significant to our data since it’s not computed rent
* Columns N to U can be combined for feature “rent\_less\_50\_percent”
* two\_or\_more\_races\_pop: redundant since we already have features with individual races
* not\_hispanic\_pop: not significant feature since we already have “Hispanic\_pop”
* other\_race\_pop: not significant feature since we are only concerned with known race
* households: can be ignored since we already have family and non-family households
* income\_per\_capita: we already have median\_income
* can be ignored since there is no data
  + pop\_5\_years\_over
  + speak\_only\_english\_at\_home
  + speak\_spanish\_at\_home
  + speak\_spanish\_at\_home\_low\_english
  + pop\_15\_and\_over
  + pop\_never\_married
  + pop\_now\_married
  + pop\_separated
  + pop\_widowed
  + pop\_divorced
* less\_one\_year\_college & one\_year\_more\_college: can be ignored since we other feature columns more important
* hispanic\_any\_race: can be ignored since not that significant to us
* commute\_35\_44\_mins, commute\_less\_10\_mins
* workers\_16\_and\_over: not significant for us
* some\_college\_and\_associates\_degree: can be ignored since we have other more relevant features
* sales\_office\_employed: can be ignored since we are looking at work from home or not scenarios
* poverty: vague classification and we already have feature columns for incomes
* mortgaged\_housing\_units: can be ignored since it has nothing to do with COVID
* Following features can be ignored since we have features for various income levels:
  + housing\_units
  + vacant\_housing\_units
  + vacant\_housing\_units\_for\_rent
  + vacant\_housing\_units\_for\_sale
  + owner\_occupied\_housing\_units
  + million\_dollar\_housing\_units
* We will ignore following features assuming we are interested into looking to family having parents or no:
  + families\_with\_young\_children
  + two\_parent\_families\_with\_young\_children
  + two\_parents\_in\_labor\_force\_families\_with\_young\_children
  + two\_parents\_father\_in\_labor\_force\_families\_with\_young\_children
  + two\_parents\_mother\_in\_labor\_force\_families\_with\_young\_children
  + two\_parents\_not\_in\_labor\_force\_families\_with\_young\_children
  + one\_parent\_families\_with\_young\_children
  + father\_one\_parent\_families\_with\_young\_children
  + father\_in\_labor\_force\_one\_parent\_families\_with\_young\_children
* We will merge following features into two features “commute\_30\_less” and “commute\_30\_more”:
  + commute\_less\_10\_mins
  + commute\_10\_14\_mins
  + commute\_15\_19\_mins
  + commute\_20\_24\_mins
  + commute\_25\_29\_mins
  + commute\_30\_34\_mins
  + commute\_35\_44\_mins
  + commute\_45\_59\_mins
  + commute\_60\_more\_mins
* aggregate\_travel\_time\_to\_work can be ignored since we are already capturing this in finer details above.
* commuters\_16\_over can be ignored because the meaning is not clear
* We will merge following features into 4 features “income\_less\_50K” , “income\_50K\_100K”, “income\_100K\_150K”, and “income\_150K\_200K:
  + income\_less\_10000
  + income\_10000\_14999
  + income\_15000\_19999
  + income\_20000\_24999
  + income\_25000\_29999
  + income\_30000\_34999
  + income\_35000\_39999
  + income\_40000\_44999
  + income\_45000\_49999
  + income\_50000\_59999
  + income\_60000\_74999
  + income\_75000\_99999
  + income\_100000\_124999
  + income\_125000\_149999
  + income\_150000\_199999
  + income\_200000\_or\_more

• We will ignore columns “CE to CW” since they are related to house built year which is not significant to our analysis

• We will ignore columns N to V since they are for the % spent on rent but we already have income level information

Analysis:

* Comparison of confirmed cases and deaths between 2020 and 2021

Prediction: